DOCUMENT RESUME

ED 427 042 TM 029 420

AUTHOR	Yi, Qing; Nering, Michael L.
TITLE	Simulating Nonmodel-Fitting Responses in a CAT Environment. ACT Research Report Series 98-10.
INSTITUTION	ACT, Inc., Iowa City, IA.
PUB DATE	1998-12-00
NOTE	37p.
AVAILABLE FROM	ACT Research Report Series, P.O. Box 168, Iowa City, IA 52243-0168.
PUB TYPE	Reports - Research (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Computer Assisted Testing; *Goodness of Fit; Item Response Theory; Models; *Responses; *Simulation; Test Coaching
IDENTIFIERS	Test Characteristic Curve

ABSTRACT

This study developed a model to simulate nonmodel-fitting responses in a computerized adaptive testing (CAT) environment, and to examine the effectiveness of the model. The underlying idea was to simulate examinees' test behaviors realistically. This study simulated a situation in which examinees are exposed to or are coached on test items before actual testing. The multidimensional item response theory (MIRT) (M. Reckase, 1985) model was adopted for this study. Test characteristic curves and the proportion of affected items administered to examinees were investigated. The results indicate that the probability of an examinee responding to an item correctly and the proportion of affected items administered to examinees were influenced by the severity and the number of affected items. The results also suggest that the proposed model might be an effective tool for investigating the issue of nonmodel-fitting responses in a CAT environment. (Contains 7 figures and 39 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ACT Research Report Series 98-10

ED 427 042

Simulating Nonmodel-Fitting Responses in a CAT Environment

Qing Yi

Michael L. Nering

TM029420

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

P. A. Farrant

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

ACT

December 1998

For additional copies write:

ACT Research Report Series
PO Box 168
Iowa City, Iowa 52243-0168

© 1998 by ACT, Inc. All rights reserved.

Simulating Nonmodel-Fitting Responses in a CAT Environment

Qing Yi
Michael L. Nering

Abstract

The purposes of this study were to develop a model to simulate nonmodel-fitting responses in a CAT environment, and to examine the effectiveness of this model. The underlying idea was to realistically simulate examinees' test behaviors. This study simulated a situation in which examinees are exposed to or are coached on test items before actual testing. The multidimensional item response theory (MIRT; Reckase, 1985) model was adopted in this study. Test characteristic curves and the proportion of affected items administered to examinees were investigated. The results indicated that the probability of an examinee responding to an item correctly and the proportion of affected items administered to examinees were influenced by the severity and the number of affected items. The results also suggested that the proposed model might be an effective tool for investigating the issue of nonmodel-fitting responses in a CAT environment.

Simulating Nonmodel-Fitting Responses in a CAT Environment

Nonmodel-fitting responses occur when an examinee responds to test items in a manner that is not congruent with the underlying test model. That is, if an examinee responds correctly to a difficult item or incorrectly to an easy item in relation to his/her ability (Reise & Due, 1991); thus, the examinee's responses are not as expected. This area of research has been known as appropriateness measurement in the past (e.g., Drasgow, 1982; Drasgow & Levine, 1986; Levine & Rubin, 1979), and as person fit more recently (e.g., Reise & Due, 1991). A variety of behaviors have been identified to explain why a person's item response pattern does not follow the underlying test model. Wright (1977), for example, suggested that examinees might be bored with a test and respond incorrectly to easy items toward the end of test. examinees might cheat on tests, or examinees might do poorly at the beginning of a test because the test format was confusing. Levine and Rubin (1979) discussed other aberrant behaviors, such as improperly aligning an answer sheet, using a poor test-taking strategy, or interpreting test questions differently from the other examinees. In 1982, Levine and Drasgow suggested the possibility of an examinee who might have high ability but due to atypical schooling or low English fluency might respond test items in a nonmodel-fitting manner.

Much of the existing research on nonmodel-fitting responses has focused on the development of statistical indices for detecting nonmodel-fitting response behaviors (e.g., Drasgow & Levine, 1986; Drasgow, Levine, & McLaughlin, 1987, Drasgow, Levine, & Williams, 1985; Levine & Rubin, 1979; Tatsuoka, 1984; van der Flier, 1982). Other

researchers have studied the properties of these statistical indices, such as the detection power (e g , Drasgow, Levine, & McLaughlin, 1987, 1991, Meijer, 1996, Nering, 1995, Reise, 1995), or the distribution properties of those indices (e g , Nering, 1995, 1996, 1997; Reise, 1995). A third research strand has been the investigation of robust ability estimation techniques to reduce the effects of nonmodel-fitting responses (e g , Meijer & Nering, 1997; Mislevy & Bock, 1982, Wainer & Wright, 1980, Yi, 1998, Yi & Nering, 1998).

Although various research has been conducted in person fit area, most of the research used Monte Carlo method. Previous research tried to develop methods to simulate nonmodel-fitting responses. However, the typical nonmodel-fitting simulation methods artificially created examinees' nonmodel-fitting response patterns that might not reflect examinees' actual test behaviors.

Typical Nonmodel-Fitting Simulation Methods

One commonly used method to simulate nonmodel-fitting response vectors is to change selected items from correct to incorrect or from incorrect to correct with certain probabilities (e g , Drasgow, Levine, & Williams,1985; Levine & Rubin, 1979). The purpose of this type of manipulation is to create either a spuriously high score (i e , changing incorrect responses to correct) that represents the situation in which a low ability examinee copies answers of difficult items from a nearby more able neighbor, or a spuriously low score (i e , switching correct responses to incorrect) in that an able examinee answers easy questions incorrectly due to language difficulties, atypical education, or alignment errors (Drasgow, Levine, & McLaughlin, 1987). Another

method to simulate nonmodel-fitting responses proposed by Reise (1995), is to reduce the value of the IRT *a* parameter (see also Meijer & Nering, 1997). These two nonmodel-fitting response simulation methods have been used in a variety of research studies (e.g., Nering, 1996). However, these types of nonmodel-fitting manipulation methods may produce artificial response patterns, rather than response patterns that may happen in real testing situations. Therefore, it is important to develop a method that may simulate examinees' nonmodel-fitting response patterns in a way that may more accurately reflect examinees' test behaviors.

Person Fit in Computerized Adaptive Testing

Most of the person-fit research has been conducted in conventional paper-and-pencil test situations. However, due to the advantages of computerized adaptive testing (CAT), the popularity of personal computers, and the vast development of computer software, CAT is now viewed as a practical alternative to traditional paper-and-pencil tests. Therefore, in order to fully understand the potential of CAT, it is necessary to study the effects of nonmodel-fitting responses in a CAT environment.

CAT is designed to overcome some of the problems encountered with traditional paper-and-pencil tests. For example, in a paper-and-pencil test every examinee is administered the same test items regardless of his/her ability level. CAT is accomplished by "tailoring" test items to an individual examinee in such a way that an examinee only receives items that appear to be appropriate for his/her estimated ability level.

The potential advantages of CAT for educational and psychological testing are well documented in the literature (e.g., Parshall, 1992; Wainer, 1990; Wise & Plake,

1989). Research has shown that an equally reliable score can be obtained in CAT with approximately half the items required in paper-and-pencil tests (e.g., McBride, 1985, Olsen, Maynes, Slawson, & Ho, 1989, Wainer, 1993, Weiss, 1982). Other advantages of CAT include frequent and convenient test scheduling, immediate scoring, computer-collection of data, and presenting items in a multimedia environment.

The measurement efficiency or shorter tests, is one often-mentioned advantage of CAT, but this efficiency may bring several practical concerns. Most computerized adaptive tests use item selection and scoring algorithms that depend on item response theory (IRT) models, which are based on strong assumptions (e.g., unidimensionality, local independence, and monitonicity). In practice, these assumptions are often violated, which may seriously compromise the quality of examinee's test scores and trait estimates. Model assumption violations could lead to a response pattern that may not fit the underlying test model, and result in an ability estimate that does not accurately reflect the latent trait of the examinee.

The problem of examinees having been coached on test items has been a particular challenging issue in real testing situations (Meijer, 1996). This problem is especially pronounced in CAT administrations because of the more frequent administration dates typically available in CAT programs (Davey & Nering, 1998). Thus, the possibility of examinees obtaining pre-knowledge of a test, that is, examinees being coached on test items or items being exposed to examinees before testing, may be increased in a CAT environment. Obtaining pre-knowledge of a test may result in exposed items becoming easier, and may result in an inflated probability of an examinee answering an item correctly.

There are three main concerns relating to test security in both paper-and-pencil and CAT administration: repeaters, cheaters, and coaching schools (Patsula & Steffen, 1997). In practice, it may be difficult to distinguish those examinees who have violated test security from those who have not. However, examinees' nonmodel-fitting responses may affect the measurement of their estimated scores.

As indicated above, there is limited research on the effects of nonmodel-fitting responses in a CAT environment. Some research has indicated that even a small number of nonmodel-fitting responses could impact the estimate of examinees' ability in a CAT environment. Nering (1996) studied the effects of nonmodel-fitting responses on test length and final $\hat{\theta}$ values within a CAT environment. The nonmodel-fitting responses were simulated by changing correct responses to incorrect or incorrect to correct responses. This study manipulated examinees' responses to 1, 2, 3, 4, or 5 items, respectively. The manipulation occurred at different points during the CAT administration: between items 1 through 5, items 3 through 7, and so forth. Nering discovered that the $\hat{\theta}$ values were inaccurate when the number of nonmodel-fitting responses increased to 4 or 5 items. He also suggested that if the nonmodel-fitting responses occurred early in the CAT administration (i.e., between items 1 and 5), the $\hat{\theta}$ values would be affected more than if the nonmodel-fitting responses occurred later in the CAT administration.

Although Nering (1996) discovered that nonmodel-fitting responses in CAT administration influenced the accuracy of ability estimates, his findings might be somewhat limited due to the methods used in the simulations. Similar to previous researchers, Nering simulated nonmodel-fitting responses simply by changing correct

responses to incorrect and incorrect to correct. Additional research is needed so that methods of simulating actual examinees' test behaviors can be developed.

Purpose

There were two purposes in this study. One was to develop a method that might realistically simulate examinees' nonmodel-fitting responses in a CAT environment. This study simulated a scenario where an examinee obtains pre-knowledge of a test (i.e., examinees have been coached on test items or they have been exposed to items before the actual testing). This type of nonmodel-fitting response could happen in a real testing situation, especially in a CAT environment. Pre-knowledge of a test may compromise the results of a test administration, thus, may provide inaccurate information about examinees' estimated ability.

The nonmodel-fitting responses were simulated in a two-stage process. In the first stage, CAT was administered to several simulees who had a relatively high ability level, and the items administered to those examinees were recorded. In the second stage of this simulation model, certain percentages of items were selected to represent the nonmodel-fitting responses based on the items administered to those high ability examinees (in stage 1). The method of simulating nonmodel-fitting responses was to reduce those selected items' item difficulty parameters. The details of the current approach of simulation are described below. The underlying idea of this simulation was to create a situation that might actually occur in a CAT administration rather than simply changing the correctness of a response.

The other purpose of this study was to evaluate this simulation model. The goal of this investigation was to determine if this newly developed simulation model functioned in a way as expected. That is, whether the number of items manipulated and the severity of the manipulation influenced the simulation results.

Method

Data Simulation

This study adopted the simulation method that was used in Yi and Nering (1998). In this study, a MIRT model was used as the template to simulate examinees' item responses (0/1s). The underlying goal was to simulate response patterns in a realistic manner (Davey, Nering, & Thompson, 1997; Parshall, Kromrey, Chason, & Yi, 1997). The simulated data were generated based on actual examinees' responses to eight forms of the ACT Mathematics Test. Sixty items were included in each test form; which resulted in a 480 item pool. Fifty-one score categories (e.g., 10, 11, 12, ..., 60; that is, the sum of the proportion correct scores in a test form) were created that represented examinees' ability level. A score category was defined to be a one unit (or more) interval (e.g., $10.5 \leq x < 11.5$, $11.5 \leq x < 12.5$, ..., where x represents a score category) on the number correct response scale. Simulees would be classified into a score category based on the sum of the P-values from the MIRT model calculation.

Procedure of CAT Administration

The high dimensional CAT model was used in this research. This high dimensional CAT model is very similar to its unidimensional counterpart (see Yi & Nering, 1998). For this study, a 25 item fixed-length CAT was implemented. The

maximum likelihood estimation (MLE) procedure was used as both the provisional and final ability estimation methods, because it is the most widely used method of ability estimation. The Sympson and Hetter (1985) item exposure control procedure was used in this study. The content balancing method that is based on the approach proposed by Davey and Thomas (1996) was implemented in the CAT administration. In this high dimensional CAT model, an examinee's item response was determined by the compensatory MIRT model. The probability of a correct response to item i in a k-dimensional compensatory normal ogive model (Davey et al., 1997) can be expressed as

$$P(u_{ij}=1|\theta_j,\mathbf{a}_i,d_i,c_i)=c_i+(1-c_i)\Phi(\mathbf{a}_i^T\theta_j+d_i)$$

where

u_{ij}	is the score (0/1) on item i $(i=1,2,3,\ldots,n)$ by person j $(j=1,2,3,\ldots,N)$,
$\mathbf{a}_i$	is the vector of item discrimination parameters $(a_{ik}=a_{i1},a_{i2},a_{i3},\ldots,a_{im})$ for item i in k dimensions $(k=1,2,3,\ldots,m)$,
d_i	is the scalar difficulty parameter for item i, negative d values represent difficult items, and positive values represent easy items,
c_i	is the scalar lower asymptote parameter for item i,
θ_j	is the vector of θ for person j,
$\Phi(\bullet)$	represents the normal distribution function, and
$P(u_{ij}=1\|\theta_j,\mathbf{a}_i,d_i,c_i)$	is the probability of an examinee j correctly answering item i

In this IRT model, there is an item discrimination parameter for each dimension of the model but only one overall item difficulty parameter. The components in the function are additive, thus, being low on one latent trait can be compensated for by being high on another trait.

Nonmodel-Fitting Responses Simulation

As indicated above, the procedure of selecting items to be nonmodel-fitting responses took two steps. In the first step, a CAT was administered to high ability simulees (i.e., $\theta = 1.5$ that was equivalent to score category around 46) whose responses probabilistically fit the underlying test model. The rationale for using high ability examinees for item selection was that these examinees are more likely to memorize items administered to them when they take a CAT, and they might then share these memorized items with other examinees. This would result in those memorized items having lower item difficulty parameter values (*d*) than when they were originally calibrated. For this stage of the study, 25 high ability examinees ($\hat{\theta} = 1.5$) were generated. In the second step of the nonmodel-fitting responses simulation, an item number matrix was developed (i.e., 25 examinees X 25 items) based on the CAT administration to this group of examinees. From this item number matrix, 24, 48, or 72 items (i.e., 5%, 10%, 15% of the 480 item pool) were randomly selected that reflected situations in which a small, medium, or large percentage of item pool was exposed. These items were then manipulated by reducing item difficulty parameter values by either 0.5, 0.75, or 1.0, representing the conditions with minor, moderate, or major affected items. The procedure of selecting items to be manipulated is displayed in Figure 1 by a flowchart.

Conditions of Study

The null condition of this study was defined as the condition in which no manipulation was done to the test items (i.e., no affected items were included). The experimental conditions, on the other hand, included the affected items, in which the percentage of item pool exposed and the severity of affected items were crossed, resulting

in 9 conditions (3 percentages of exposed item pool X 3 levels of severity of affected items). A total of 10 study conditions were involved in the current research (i.e., 1 null and 9 experimental conditions).

Data Analysis

The test characteristic curves (TCCs) and the proportion of affected items administered to examinees under various study conditions were used as ways to evaluate the effectiveness of this nonmodel-fitting response simulation model. To determine how the affected items of an item pool influenced the resulting TCCs, the average *P*-values under different study conditions were examined conditionally on the score category. This was done as a way of evaluating how the percentage of exposed item pool and the severity of affected items influenced the calculation of average *P*-values along the ability continuum. TCCs were obtained as the sum of the *P*-values along the score category. The *P*-values were calculated based on the MIRT model on those selected items. The TCCs for the experimental conditions were compared to those found in the null condition.

The proportion of affected items administered to examinees was also studied. The goal was to determine if the percentage of exposed item pool and the severity of affected items in fact impacted whether affected items would be administered to examinees. In addition, this investigation also might provide information in terms of what kind examinee (e.g., at which point on the score category) would be affected the most by gaining pre-knowledge of a test.

Results

Visual display was used as the method to summarize the results for this study. This method has the advantage of easily displaying the differences among conditions. Figures 2 to 4 present the TCCs under different experimental conditions (e.g., 5%, 10%, or 15% exposed item pool and 0.5, 0.75, or 1.0 reduced item difficulty parameter values). The TCCs from the null conditions were compared to the TCCs from the experimental conditions. The calculations of the average *P*-values were based on the situation as if there were 24, 48, or 72 items in the item pool and all those items were exposed to examinees with different levels of affected items before the actual testing. Figure 2 displays the TCCs under the 5% (i.e., 24 items) exposed item pool condition across the four levels of affected items (i.e., no, minor, moderate, and major affected items). As the levels of severity of affected items increased, so did the difference between average *P*-values from the experimental and the null conditions (see Figure 2). The average *P*-values appeared to be higher in the experimental conditions than those in the null condition. Examinees with abilities at the extreme score categories were not affected as much as examinees with mid-range score categories.

Figures 3 and 4 present the TCCs under the 10% (i.e., 48 items) and the 15% (i.e., 72 items) exposed item pool conditions. The average *P*-values obtained under the experimental conditions were larger than those from the null condition, especially when the nonmodel fitting became more severe. It seemed that the severity of affected items influenced the TCCs more than the percentage of exposed item pool. Across Figures 2 to 4, it is clear that the patterns of the TCCs shifting from the null conditions were similar

under these three percentages of exposed item pool conditions, but the amount of shifting increased as the severity of affected items became higher

The results obtained from the calculation of the average *P*-values indicated that the model for simulating the nonmodel-fitting responses that was developed in this study affected the resulting TCCs Average *P*-values shifted to higher values when there were nonmodel-fitting responses in comparison to the average *P*-values under the null conditions This shift increased when the nonmodel fitting became more severe (e.g., larger percentage of exposed item pool and major affected items)

Figures 5 to 7 present the results of the proportion of affected items administered to examinees at each score category under different study conditions. The proportion of affected items administered to examinees under the 5% exposed item pool (i.e., 24 items) conditions across the three levels of severity of affected items is displayed in Figure 5 The shapes of the proportion of affected items administered to examinees were relatively similar across these three levels of severity of affected items (i.e., minor, moderate, and major affected items) There were fewer affected items administered to examinees at extreme ability levels (e.g., low or high score categories) The maximum proportion of affected items administered to examinees was about 0.27 (i.e., about 7 affected items administered to examinees out of a 25 item test) at the score category around 46 Thus, examinees whose estimated ability was close to the point ($\hat{\theta} = 1.5$, around score category 46) from where nonmodel-fitting items were selected were administered the maximum proportion of affected items

Figure 6 displays the proportion of affected items administered to examinees under the 10% (i.e., 48 items) exposed item pool conditions Similar patterns were

discovered in the 10% exposed item pool conditions as those in 5% exposed item pool conditions. The influence of the severity of affected items on the proportion of affected items administered to examinees was relatively small. The maximum proportion of affected items given to examinees was around score category 46. However, when there were 10% items exposed, the maximum percentage of affected items administered to examinees increased to about 63% (i.e., about 16 out of 25 administered items were affected items) of the total administered items. Unsurprisingly, there were more affected items administered to examinees around the point (i.e., around score category 46 or θ approximately equaled to 1.5) from where the affected items were originally selected.

Figure 7 presents the results obtained from the 15% (i.e., 72 items) exposed item pool conditions. Similar to the 5% and the 10% exposed item pool conditions, the severity of affected items did not have big effects on the number of affected items administered to examinees. However, the maximum percentage of affected items administered to examinees increased to about 69% under the 15% exposed item pool conditions (i.e., about 17 out of 25 item test were affected items). Examinees whose estimated ability close to 1.5 were administered the maximum percentage of affected items.

The results of the proportion of affected items administered to examinees indicated that examinees whose estimated ability close to the ability from where the affected items were selected were influenced the most by the pre-knowledge of a test. In addition, the proportion of affected items administered to examinees also was affected by the percentage of exposed item pool. The larger the percentage of affected item pool, the more affected items administered to examinees. The severity levels of affected items, on

the other hand, had relatively small effects on the proportion of affected items administered to examinees.

Discussion

CAT has efficiency as an often-mentioned advantage; however, this efficiency may cause practical concerns in real testing situations particularly when test security has been compromised. For example, if examinees have been coached on the test or they have been exposed to items before testing, then, the reliability and the validity of the test results will be questionable. There has been limited research on the influence of nonmodel-fitting responses in a CAT environment. In addition, typical methods of simulating nonmodel-fitting responses do not reflect actual examinees' test behaviors (e.g., Levine & Rubin, 1979). The main goal of this study was to develop a more realistic nonmodel-fitting response simulation model in CAT environment.

The results of this study indicated that the proposed model for simulating the nonmodel-fitting responses influenced the calculation of the average *P*-values and the number of manipulated items administered to examinees. The change in the average *P*-values for examinees was influenced by the severity of the reduction in item difficulty parameter. The estimated *P*-values were higher for examinees responding to items in a nonmodel-fitting way.

The proportion of affected items administered to examinees increased as the percentage of exposed item pool increased. The maximum percentage of affected items administered to examinees occurred around score category 46, that is, the point at which the affected items were initially selected. The results indicated that if a certain part of an

item pool was exposed to examinees who had relative high ability, then other examinees who had similar abilities would have more chance to be administered these exposed items.

This study was an initial investigation of the effects of nonmodel-fitting responses in a CAT environment. The nonmodel-fitting simulation model appeared to function in an expected manner; however, additional research is needed to further study the effects of nonmodel-fitting responses in CAT administration. This study only simulated one kind of nonmodel-fitting behavior (i.e., examinees obtaining knowledge of test items before actual testing). Models that simulate other nonmodel-fitting responses need to be developed, such as cheating behavior, or distracted behavior in a test administration. In practice, it is difficult to differentiate between nonmodel-fitting and model-fitting responses; however, it is important to provide accurate information about examinees' test performance. Therefore, future research needs to investigate whether certain CAT procedures (e.g., robust ability estimation methods or robust item selection methods) can alleviate the effects of nonmodel-fitting responses, and can provide accurate measurement on assessment tasks.

References

Chang, H. & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20,* 213-229.

Davey, T. & Nering, M. (1998, September). *Controlling item exposure and maintaining item security.* Paper presented at the Computer-Based Testing: Building the Foundation for Future Assessments Colloquium Sponsored by Educational Testing Service.

Davey, T., Nering, M, & Thompson, T. (1997, June). *Realistic simulation procedures for item response data.* In T. Miller (Chair), *High-dimensional simulation of item response data for CAT research.* Symposium conducted at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Davey, T. & Thomas, L. (1996, April). *Constructing adaptive tests to parallel conventional programs.* Paper presented at the annual meeting of American Educational Research Association, New York.

Drasgow, F. (1982). Choice of test model for appropriateness measurement. *Applied Psychological Measurement, 6,* 297-308.

Drasgow, F. & Levine, M. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10,* 59-67.

Drasgow, F., Levine, M., & McLaughlin, M. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11,* 59-79.

Drasgow, F., Levine, M., & McLaughlin, M. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15,* 171-191.

Drasgow, F., Levine, M., & Williams, E. (1985). Appropriateness measurement with ploychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86.

Levine, M. & Drasgow, F. (1982). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement, 43,* 675-685.

Levine, M. & Rubin, D. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4,* 269-290.

McBride, J. (1985). Computerized adaptive testing. *Educational Leadership, 43,* 25-28.

Meijer, R. (1996). The influence of the presence of deviant item score patterns on the power of a person-fit statistic. *Applied Psychological Measurement, 20,* 141-154.

Meijer, R., & Nering, M. (1997). Trait level estimation for nonfitting-response vectors. *Applied Psychological Measurement, 21(4),* 321-336.

Mislevy, R. J., & Bock, R. D. (1982). Biweight estimators of latent ability. *Educational and Psychological Measurement, 42,* 725-737.

Nering, M. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19,* 121-129.

Nering, M. (1996). *The effects of person misfit in computerized adaptive testing.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Nering, M. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21,* 115-127.

Nering, M, Thompson, T., & Davey, T. (1997, June). *Simulation of realistic ability vectors.* In T. Miller (Chair), *High-dimensional simulation of item response data for CAT research.* Symposium conducted at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Olsen, J., Maynes, D., Slawson, D., & Ho, K. (1989). Comparisons of paper-administered, computer-administered and computerized adaptive achievement tests. *Journal of Educational Computing Research, 5,* 311-326.

Parshall, C. G. (1992). *Computer testing vs. paper-and-pencil testing: An analysis of examinee characteristics associated with mode effects on the GRE General Test.* Unpublished doctoral dissertation. University of South Florida.

Parshall, C. G., Kromrey, J. D., & Chason, W. M., & Yi, Q. (1997, June). *Evaluation of parameter estimation under modified IRT models and small samples.* Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Patsula, L. N. & Steffen, M. (1997, March). *Maintaining item and test security in a CAT environment: A simulation study.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Reckase, M. (1985, April). *The difficulty of test items that measure more than one ability.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Reckase, M., Thompson, T., & Nering, M. (1997, June). *Identifying similar item content clusters on multiple test forms.* In T. Miller (Chair), *High-dimensional simulation*

of item response data for CAT research. Symposium conducted at the annual meeting of the Psychometric Society, Gatlinburg, TN

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19(3),* 213-229.

Reise, S. P. & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15,* 217-226.

Sympson, J. B. & Hetter, R. D. (1985*). Controlling item-exposure rates in computerized adaptive testing.* Proceedings of the 27th annual meeting of the Military Testing Association (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Tatsuoka, K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 95-110.

Thompson, T., Nering, M., & Davey, T. (1997, June). Multidimensional IRT scale linking without common items or common examinees. In T. Miller (Chair), *High-dimensional simulation of item response data for CAT research.* Symposium conducted at the annual meeting of the Psychometric Society, Gatlinburg, TN.

van der Flier, W. H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology, 13,* 267-298.

Wainer, H. (1990). *Computerized adaptive testing: A primer.* Hillsdale NJ: Erlbaum.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice, 12(1),* 15-20.

Wainer, H. & Wright, B. (1980). Robust estimation of ability in the Rasch model. *Psychometrika, 45,* 373-391.

Weiss, D. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473-492.

Wise, S. & Plake, B. (1989). Research on the effects of administering tests via computers. *Educational Measurement: Issues and Practice, 3,* 5-10.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 95-115.

Yi, Q. & Nering, M. (1998, April). *Nonmodel-fitting responses and robust ability estimation in a realistic CAT environment.* Paper presented at the annual meeting of American Educational Research Association, San Diego, CA.

Yi, Q. (1998). *A comparison of three ability estimation procedures for computerized adaptive testing in the presence of nonmodel-fitting responses resulting from a comprised item pool.* Unpublished doctoral dissertation, University of South Florida, Tampa.

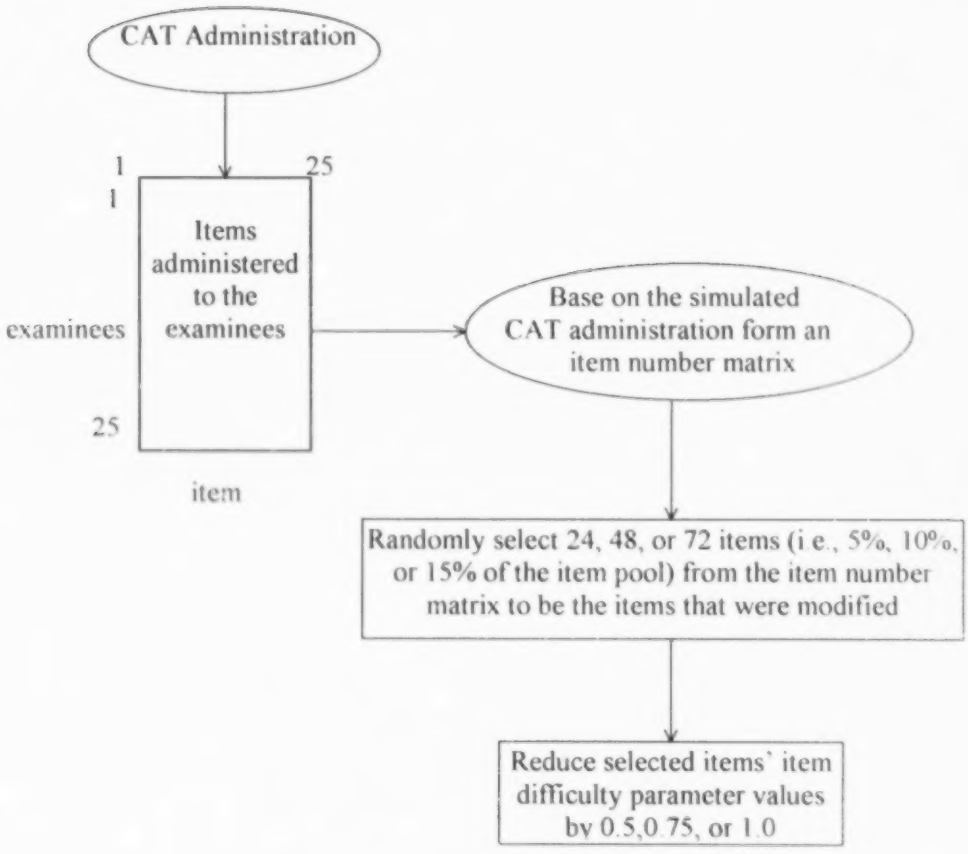
CAT Administration
1
25
1
Items administered to the examinees
examinees
25
item
Base on the simulated CAT administration form an item number matrix
Randomly select 24, 48, or 72 items (i e., 5%, 10%, or 15% of the item pool) from the item number matrix to be the items that were modified
Reduce selected items' item difficulty parameter values by 0.5,0.75, or 1.0

FIGURE 1. Procedure of Selecting Items to be Affected Items

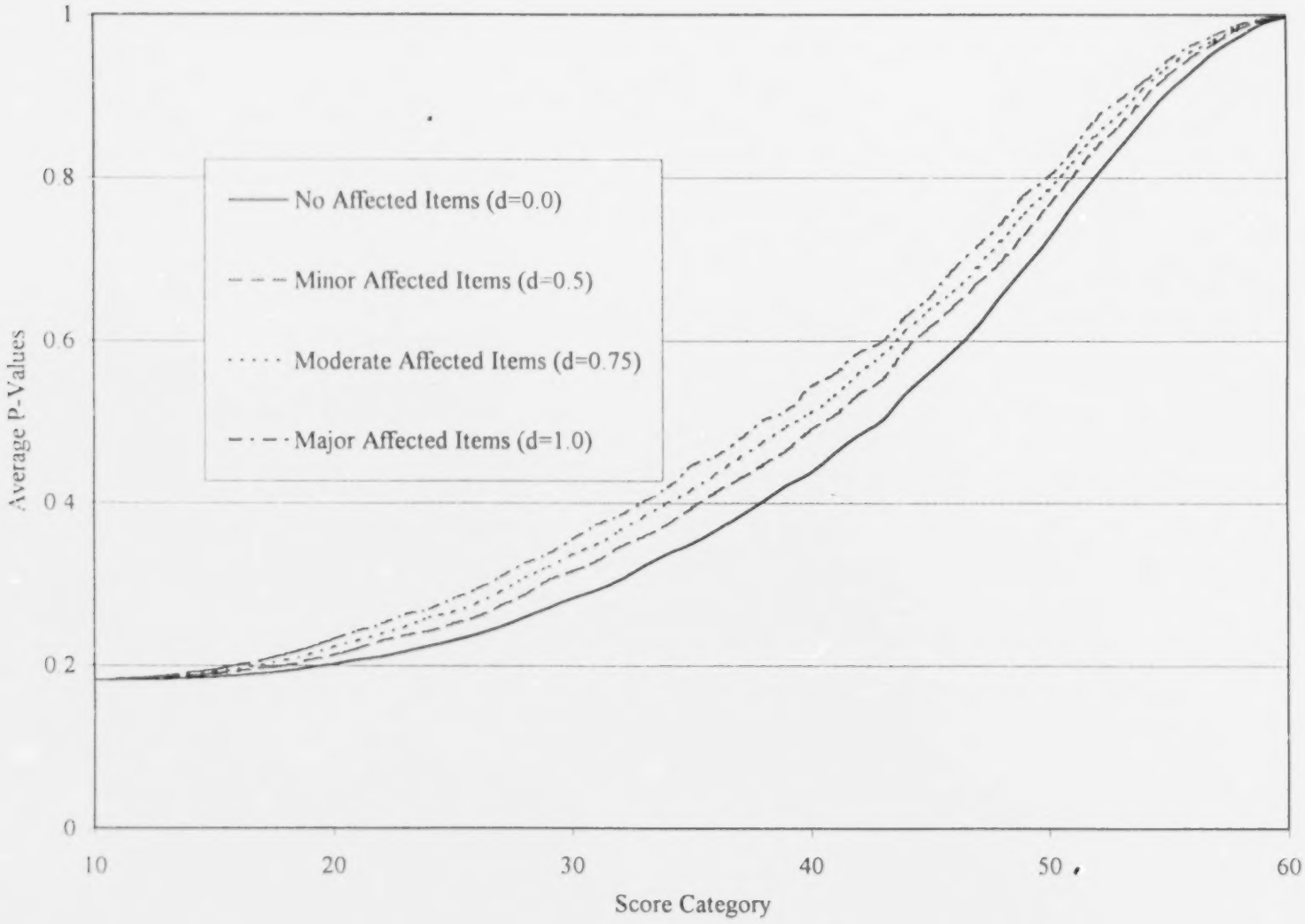

FIGURE 2. Test Characteristic Curves Under Small Percentage of Exposed Item Pool Condition (5%) Across Four Levels of Affected Items

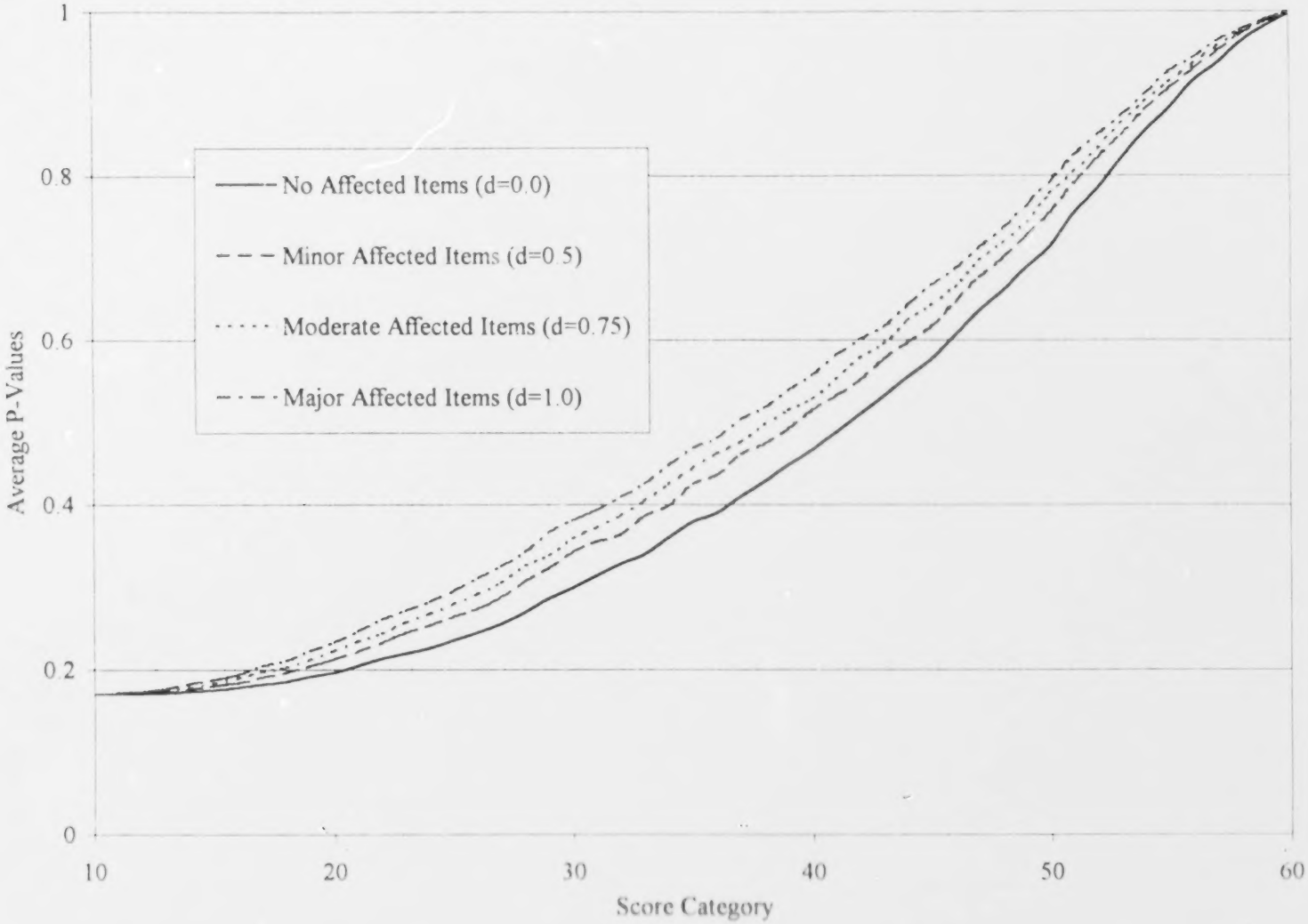

FIGURE 3. Test Characteristic Curves Under Medium Percentage of Exposed Item Pool Condition (10%) Across Four Levels of Affected Items

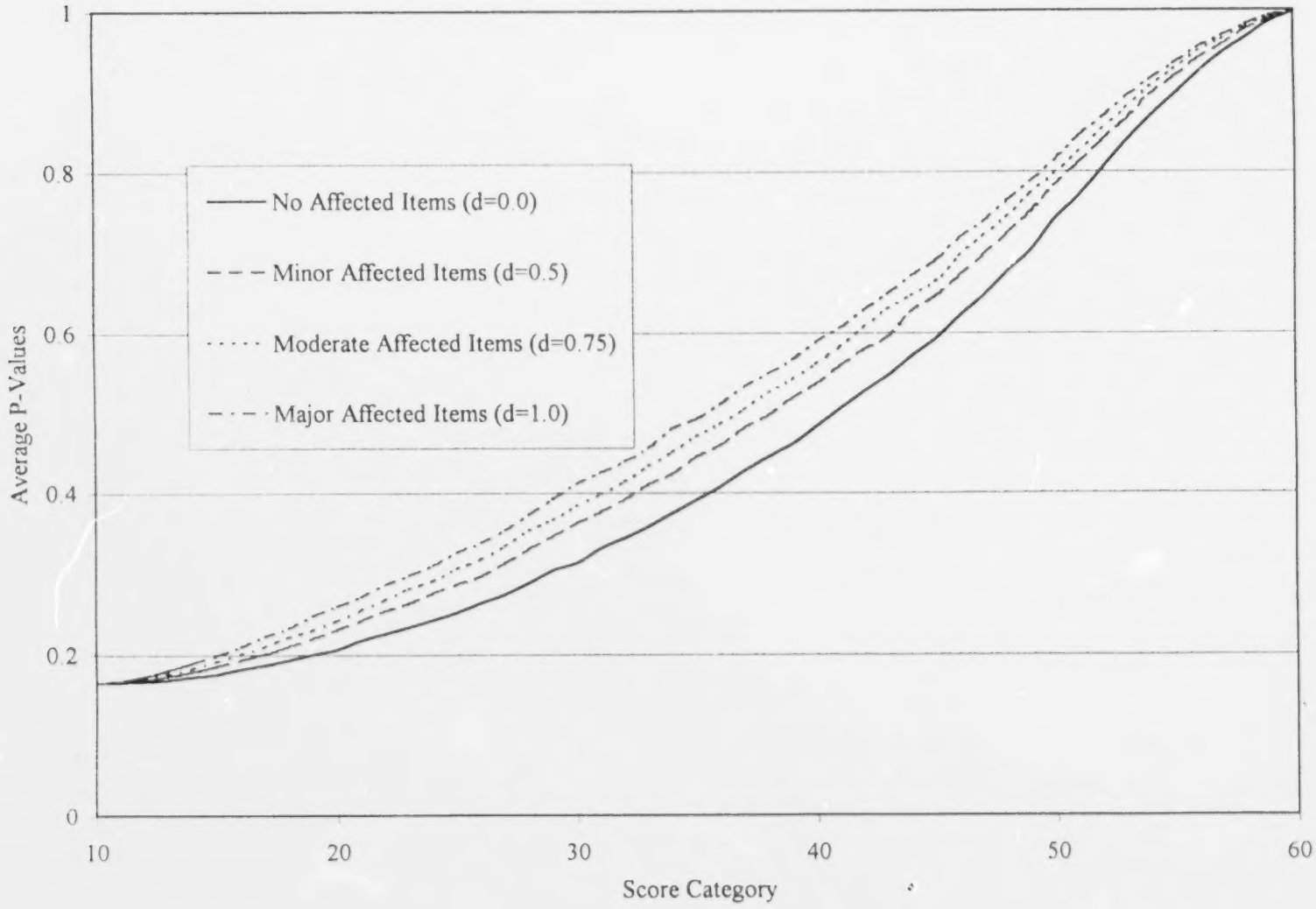

FIGURE 4. Test Characteristic Curves Under Large Percentage of Exposed Item Pool Condition (15%) Across Four Levels of Affected Items

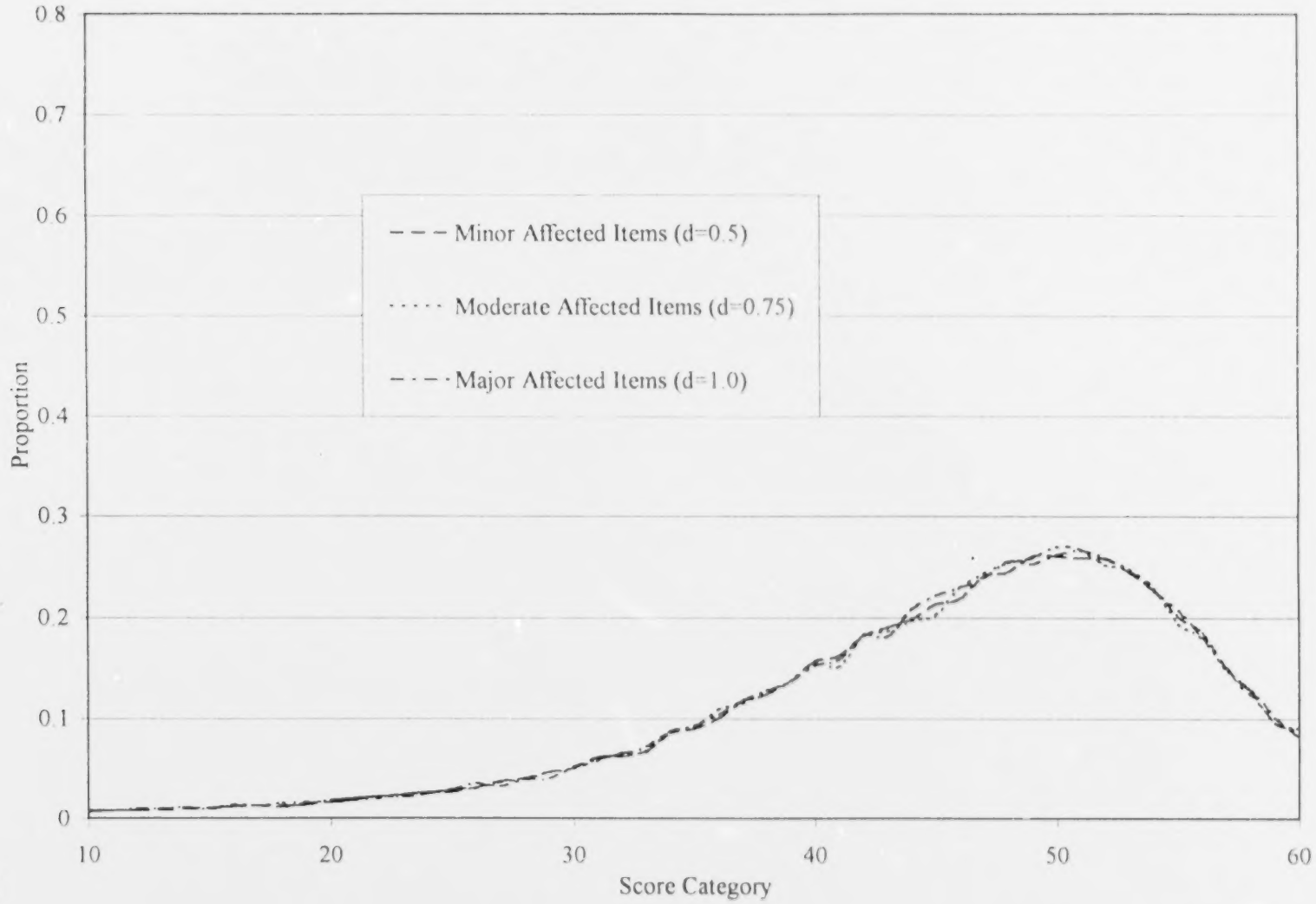

FIGURE 5. Proportion of Affected Items Administered to Examinees Under Small Percentage of Exposed Item Pool (5%) Across Three Levels of Affected Items

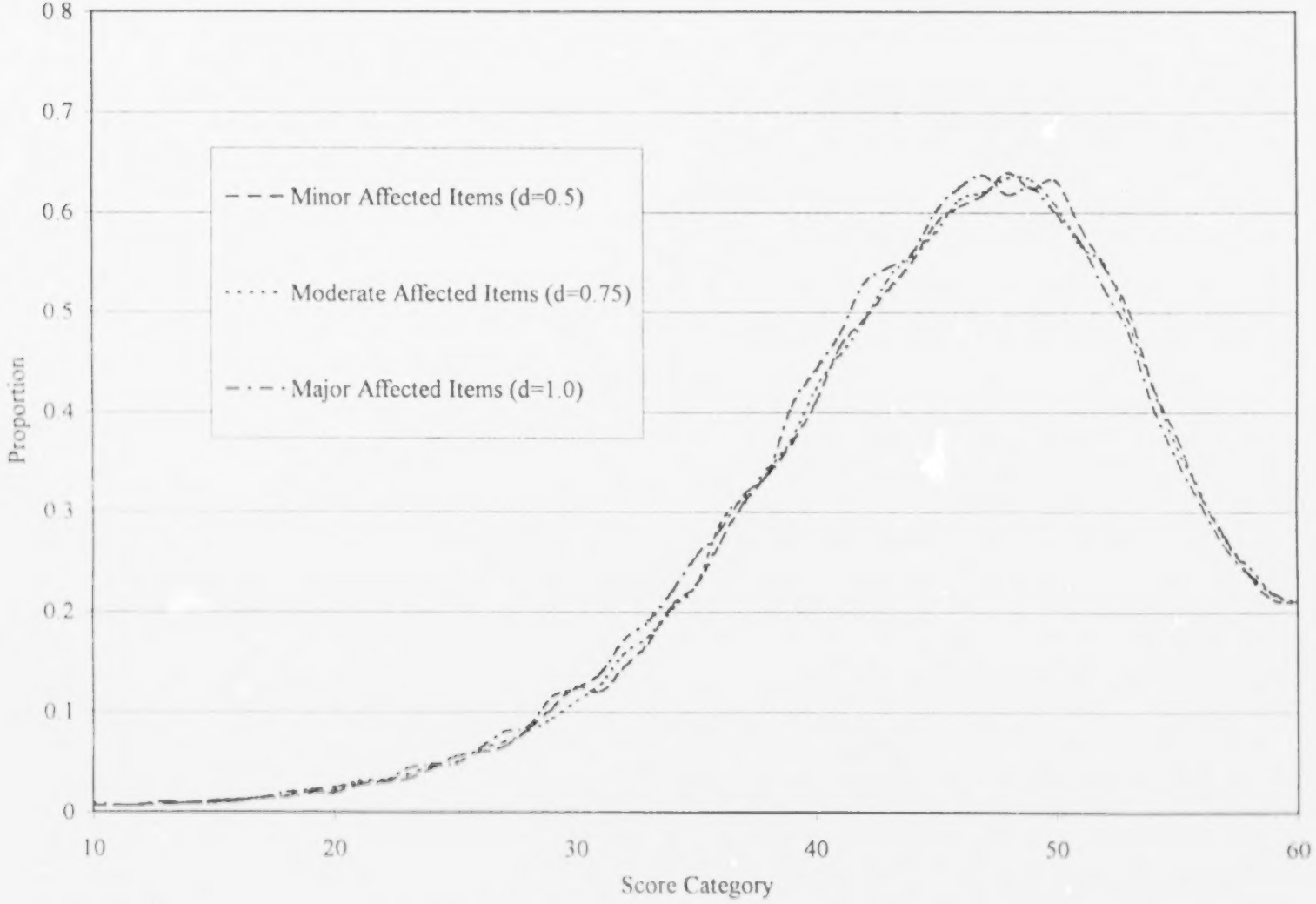

FIGURE 6. Proportion of Affected Items Administered to Examinees Under Medium Percentage of Exposed Item Pool (10%) Across Three Levels of Affected Items

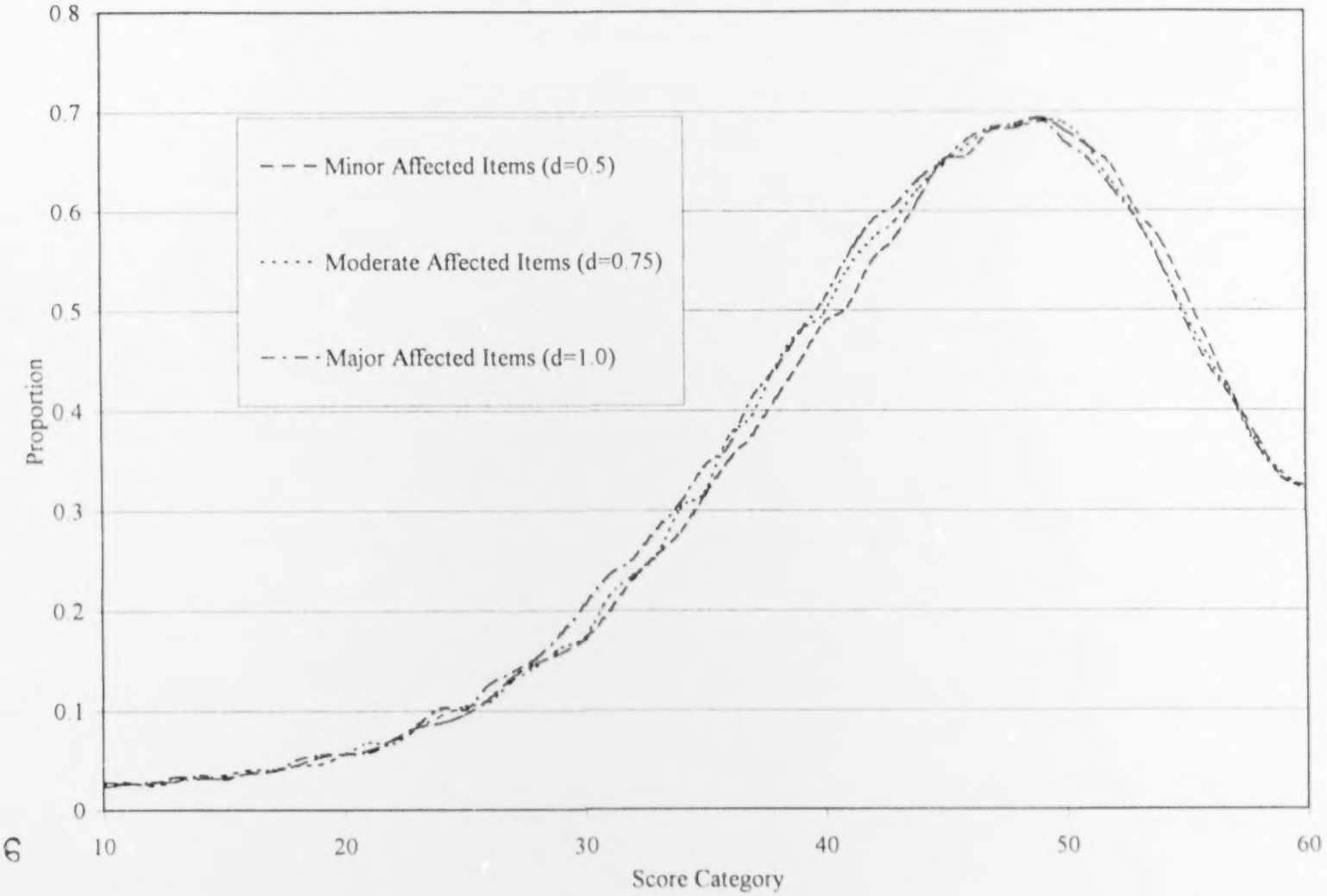

FIGURE 7. Proportion of Affected Items Administered to Examinees Under Large Percentage of Exposed Item Pool (15%) Across Three Levels of Affected Items